(51) International Patent Classification 6 : A61B 5/00	A1	(11) International Publication Number: $\quad$ WO 98/35609 (43) International Publication Date: $\quad$ 20 August 1998 (20.08.98)

(54) Title: A SYSTEM FOR PREDICTING FUTURE HEALTH

(57) Abstract

A computer–based system is disclosed for predicting future health of individuals comprising: (a) a computer comprising a processor containing a database of longitudinally–acquired biomarker values from individual members of a test population, subpopulation D of said members being identified as having acquired a specified biological condition within a specified time period or age interval and a subpopulation $\overline{D}$ being identified as not having acquired the specified biological condition within the specified time period or age interval; and (b) a computer program that includes steps for: (1) selecting from said biomarkers a subset of biomarkers for discriminating between members belonging to the subpopulations D and $\overline{D}$, wherein the subset of biomarkers is selected based on distributions of the biomarker values of the individual members of the test population; and (2) using the distributions of the selected biomarkers to develop a statistical procedure that is capable of being used for: (i) classifying members of the test population as belonging within a subpopulation PD having a prescribed high probability of acquiring the specified biological condition within the specified time period or age interval or as belonging within a subpopulation $\overline{PD}$ having a prescribed low probability of acquiring the specified biological condition within the specified time period or age interval; or (ii) estimating quantitatively, for each member of the test population, the probability of acquiring the specified biological condition within the specified time period or age interval.

A SYSTEM FOR PREDICTING FUTURE HEALTH

FIELD OF INVENTION

A computer-based system and method are disclosed for predicting the future health of an individual. More particularly, the present invention predicts the future health of an individual by obtaining longitudinal data for a large number of biomarkers from a large human test population, statistically selecting predictive biomarkers, and determining and assessing an appropriate multivariate evaluation function based upon the selected biomarkers.

BACKGROUND OF THE INVENTION

It would be desirable if the onset of future health problems could be predicted for an individual with sufficient reliability far enough into the future so that the chances could be increased for preventing future health problems for that individual rather than waiting for actual onset of a disease and then treating the symptoms. At present, the overwhelming fraction of medical research funding is directed toward improving methods of diagnosis and treatment of disease rather than toward discovering preventive measures that could be directed toward reducing the risk of disease long before any of the typically observed symptoms of the disease are evident. Although the emphasis on treatment of diseases may have led to enormous advances in the medical sciences in terms of the large number and great sophistication of the techniques and methods developed for diagnosing existing diseases as well as for treating the diseases after diagnosis, such advances continue to lead to ever-increasing costs for treatment. Such costs can have staggering financial consequences for individuals as well as for the entire society. Such staggering costs have led to increasing public pressure to find ways of reducing medical costs.

Thus, in addition to the benefit to be gained by an individual who could be informed of the high risk of the onset of disease far enough in advance so that effective preventive steps could be taken, substantial reductions in overall medical costs might be realized by entire communities and/or countries.

A living person will be selected at random from all U.S. residents and followed for a period of one year. At the end of the year the person's vital status (alive or dead) will be ascertained. The "event" is "the person died during the year." At the end of the year the event either occurred (person died) or did not occur (person survived) with *post hoc* probabilities of 1 and 0, respectively. Before the person is selected, the U.S. mortality statistics can be used to estimate the *a priori* probability that the person will die in the year. This probability is computed as $p=d/N$, where N is the total number of persons in the *at risk* group (here, all the persons in the U.S. population who were alive at the beginning of the year) and d is the total number of deaths among the at risk group. For example, the data from calendar year 1993 are (approximately), $d = 2,268,000$, $N = 257,932,000$, and the *a priori* probability of the event is approximately $p = 0.0088$. [Data from *Microsoft Bookshelf 1995 Almanac*, article entitled, "Vital Statistics, Annual Report for the Year 1993 (Provisional Statistics), Deaths." and *Vital Statistics of the United States*, published by the National Center for Health Statistics.] In this game, the *a priori* probability of the event is based upon very little information, simply that the person would be a member of the at risk group, consisting of all persons who would be alive and a U.S. resident at the time of selection.

Additional information about the at risk group, from which the subject is selected at random, implies additional information about the subject and modification of the *a priori* probability of the event. For example, continuing the "game" above, based on 1993 data:

- If the at risk group were the group of U.S. males, *i.e.*, if the subject is known, *prior to selection*, to be a male, the *a priori* probability of the event is approximately $p = 0.0093$, which is about 6% higher than the case where gender is unknown or unspecified.

- If the at risk group were the group of U.S. males aged 75-84, *i.e.*, if the subject is known, *prior to selection*, to be a male in the age interval 75-84, the *a priori* probability of the event is approximately $p = 0.0772$, or about 8.3 times as high as for males where age is unknown or unspecified.

statistical definition of *risk* as expected loss, where the loss function takes value 1 if the event occurs and 0 if the event does not occur.

The foregoing comments illustrate the principle that differing levels of information lead to differing *a priori* probabilities. The risk for a person about whom much is known (*i.e.*, a member of a small subpopulation with many known characteristics) may be very different from the risk for a large subpopulation with few known characteristics. However, there is yet another problem confounding the ability of traditional scientific research studies on populations to ascertain risk of disease for individuals. This problem results from a commonly over-simplified understanding of the causation of disease, particularly the causation of chronic degenerative diseases such as cancers, cardiovascular diseases, diabetes, etc. That is, there is a tendency to believe, for a variety of reasons, that such diseases can either be controlled or be clinically indicated by single constituents or by prescribing a single pharmaceutical compound. For example, it has been suggested that breast cancer can be controlled by a modest reduction of fat intake, that colon cancer can be controlled by adding specific dietary fiber components, that heart disease is clinically indicated by elevated blood cholesterol, and that stomach cancer can be clinically indicated by low blood levels of vitamin C. These over-simplified views too often prove to be inadequate for identifying causation, particularly for an individual person. There are too many confounding variables to be taken into consideration, to say nothing of the great difficulties of extrapolating population data to individuals within the population. Testing and investigating single constituents, among a milieu of thousands if not millions of possible constituent causes, is fraught with great uncertainty, especially when attempting to extrapolate these data to the estimation of disease risks for individuals.

These dual difficulties, (a) of extrapolating data for experimental populations of individuals to a randomly selected individual and (b), of relying on single indicators or causes of disease occurrence, seriously compromise estimation of future disease risk for a randomly selected individual. If an individual's risk for a specific disease could be determined more reliably, it then would be possible to provide information to this individual who could then make more informed decisions on his or her personal behavior. In essence, much more reliable methods

probability that an individual will acquire a specified biological condition within a specified time period or age interval and uses cross-sectional and/or longitudinal values of those biomarkers to estimate the individual's risk.

5   Still more particularly, the present invention is directed to a computer-based system for predicting future health of individuals comprising:

(a) a computer comprising a processor containing a database of longitudinally-acquired biomarker values from individual members of a test population, subpopulation D of said members being identified as having acquired a specified biological condition within a

10   specified time period or age interval and a subpopulation $\bar{D}$ being identified as not having acquired the specified biological condition within the specified time period or age interval; and

(b) a computer program that incudes steps for:

(1) selecting from said biomarkers a subset of biomarkers for discriminating

15   between members belonging to the subpopulations D and $\bar{D}$, wherein the subset of biomarkers is selected based on distributions of the biomarker values of the individual members of the test population; and

(2) using the distributions of the selected biomarkers to develop a statistical procedure that is capable of being used for:

20   (i) classifying members of the test population as belonging within a subpopulation PD having a prescribed high probability of acquiring the specified biological condition within the specified time period or age interval or as belonging within a subpopulation $\overline{PD}$ having a prescribed low probability of acquiring the specified biological condition within the specified time period or age interval; or

25   (ii) estimating quantitatively, for each member of the test population, the probability of acquiring the specified biological condition within the specified time period or age interval.

The present invention is further directed, *inter alia*, to a computer-based system for predicting

30   an individual's future health comprising:

(a) a computer comprising a processor containing a plurality of biomark... ...es

7

Example.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention will now be described in detail for specific preferred embodiments of
5   the invention, it being understood that these embodiments are intended as illustrative
examples and the invention is not to be limited thereto.

The present invention is based on the theory that an individual's health is, in general,
influenced by a complex interaction of a wide range of physiological and biochemical
10   parameters relating to the nutritional, toxicological, genetic, hormonal, viral, infective,
anthropometric, lifestyle and any other states potentially describing the aberrant physiological
and putative pathological states of that individual. Based on this theory, the present invention
is directed towards providing a practical system for predicting future health using multivariate
statistical analysis techniques that are capable of providing quantitative predictions of one's
15   future health based on statistically comparing an individual's set of biomarker values with a
longitudinally-obtained database of sets of a large number of individual biomarker values for
a large test population. The term "biomarker" is used herein to refer to any biological
indicator that may affect or be related to diagnosing or predicting an individual's health. The
term "longitudinal" is used herein to refer to the fact that the biomarker values are to be
20   periodically obtained over a period of time, in particular, on at least two measurement
occasions.

The frequency and duration of longitudinal assessments may vary. For example, some
biomarkers may be assessed annually, for periods ranging from as short as 2 years to a period
25   as long as a total lifetime. Under some circumstances, such as evaluation of newborn
children, biomarkers could be assessed more frequently as, for example, daily, weekly, or
monthly. Longitudinal assessment occasions may be "irregularly timed," *i.e.*, occur at
unequal time intervals. The set of longitudinal assessments for an individual may be
"complete," meaning that data from all scheduled assessments and all scheduled biomarkers
30   are actually obtained and available, or "incomplete," meaning that the data are not complete
in some manner. An individual's biomarkers may be assessed either cross-sectionally, *i.e.*, at

9

Phase II is a Parameter Estimation Phase that uses mixed linear models to estimate expected value vector and structured covariance matrix parameters of the Candidate Biomarkers, even in the presence of incomplete data and/or irregularly timed longitudinal data. Phase III is a Biomarker Selection and Risk Assessment Phase that uses discriminant analysis methodology and logistic regression to select informative biomarkers (including, where relevant, longitudinal assessments), to estimate discriminant function coefficients, and to use an inverse cumulative distribution function and logistic regression to estimate each individual's risk. Phase IV is an Evaluation Phase that uses the Evaluation Sample to produce unbiased estimates to the misclassification rates of the discriminant procedure.

Although the individual steps of the statistical procedures noted in the previous paragraph are described in the statistical literature, it is believed that these individual steps have never been combined in a single overall procedure as disclosed herein. In particular, classical versions of the following procedures are described for example, in the Encyclopedia of Statistical Sciences, edited by Samuel Kotz, Normal L. Johnson, and Campbell B. Read, published by John Wiley & Sons, 1985 and in additional literature cited therein: (a) correlation analysis (Volume 2, pp. 193-204), (b) logistic regression analysis (Volume 5, pp. 128-133), (c) mixed model analysis (Volume 3, pp. 137-141, article "Fixed-, Random-, and Mixed-Effect Models"), (d) discriminant analysis (Volume 2, pp. 389-397). The present invention can utilize classical versions of these procedures or such enhancements to and newer versions of these procedures as may be developed and published from time to time.

Correlation analysis is a term for statistical methods used for estimating the strength of the linear relationship between two or more variables. Correlation, as used here, can include a variety of types of correlation, including but not limited to: Pearson product-moment correlations, Spearman's ρ, Kendall's τ, the Fisher-Yates r_F, and others.

Logistic regression is a term for statistical methods, including log-linear models, used for the analysis of a relationship between an observed dependent variable (that may be a proportion, or a rate) and a set of explanatory variables. The applications of the logistic regression (or other log-linear models) used herein are primarily for the analysis in which the dependent

function that is used as the basis for calculating an estimate of the probability that a given observation belongs in a given group. For the present invention, the observations of interest typically comprise a plurality of biomarker values that are obtained from each member of a large test population or from an individual test subject. The discriminant functions of the

5   present invention are developed using distributions of these biomarker values for each biomarker determined to be of interest. Such distributions plot the total number of individual members of the test population having each biomarker value vs. the biomarker value itself. Thus, the present invention empolys a statistical procedure that uses distributions based on the individual biomarker values that are obtained for each biomarker from individual

10   members from the test population, as distinct, for example, from using mean biomarker values that are obtained from different test populations for the different biomarkers.

The term "discriminant function" is intended to mean any one of several different types of functions or procedures for classifying an observation (scalar or vector) into two or more

15   groups, including, but not limited to, linear discriminant functions, quadratic discriminant functions, nonlinear discriminant functions, and various types of so-called optimal discriminant procedures.

The computer-based system of the present invention includes a computer comprised of a

20   processor that is capable of running a computer program or set of computer programs (hereinafter refined to simply as "the computer program") comprising the steps for performing the required computations and data processing in the various steps and phases of the present invention. The processor may be a microprocessor, a personal computer, a mainframe computer, or in general, any digital computer that is capable of running computer

25   programs that can perform the required computations and data processing. The processor typically includes a central processing unit, a random access memory (RAM), read-only memory (ROM), one or more buses or channels for transfer of data among its various components, one or more display devices (such as a "monitor"), one or more input-output devices (such as floppy disk drives, fixed disk drives, printers, etc.), and adapters for

30   controlling input-output devices and/or display devices and/or connecting such devices to the buses/channels. A particular processor may include all of these components or only a subset

13

definable and measurable biomarkers.

As one of the unique features of the present invention, the subject computer-based system and apparatus may be used to determine the risk of a specified individual acquiring any one of these major diseases based on comparing that individual's profile of biomarker values with the biomarker values obtained from members of a large test population. Since it is known that these major diseases share many common factors that may be reflected in the biomarker values, the subject computer-based system may be used to concurrently assess the risk of acquiring any of these major diseases. For example, it is known that total serum cholesterol is a biomarker that is related to many of these diseases. By monitoring each profile of biomarker values that is a significant predictor, in combination with other significant biomarker predictors, of a specific disease or underlying cause of death and using the present invention to compare that profile with the test populations, an individual subject may be informed, with specified quantitative reliability, which disease poses the greatest risk for that specific individual.

A particular feature of the present invention is that those individuals who are at greatest risk of acquiring a specified disease may be provided with a quantitative probability of acquiring that disease within a specified time period or age interval in the future well before any of the typical symptoms of that disease are manifest. Armed with that information, for the many diseases known to be responsive to altered dietary and lifestyle conditions, that individual may then make those behavioral changes that can reduce the risk of the disease identified.

Furthermore, as more and more data are acquired for larger and larger numbers of subjects over longer and longer periods of time, more and more refined divisions of each of the major diseases and causes of deaths as well as of the less common diseases and underlying causes of death can be defined and included in the methodology of the present invention. For example, a breakdown can be made in terms of the different types of cancer, e.g., liver cancer, lung cancer, stomach cancer, prostate cancer, etc. The present computer-based system, thus, provides a means for including ever larger fractions of the population, so as to predict the quantitative risk of each individual acquiring, or not acquiring, a specified pathologically

preferably collected and recorded for each member of the test population each time a biological sample is taken.

TABLE 1. An illustrative list of biomarkers that may be used in the subject method for predicting future health.

SERUM BIOMARKERS
Total cholesterol*
HDL cholesterol*
10 LDL cholesterol*
Apolipoprotein b*
Apolipoprotein A_1*
Triglycerides*
Lipid peroxide (Malondialdehyde
15 equivalency:TBA)*
α- Carotene (corrected for lipoprotein carrier)*
β-Carotene (corrected for lipoprotein carrier)*
20 γ-Carotene (corrected for lipoprotein carrier)*
zeta-Carotene (corrected for lipoprotein carrier)*
α-Cryptoxanthin (corrected for lipoprotein
25 carrier)*
β-Cryptoxanthin (corrected for lipoprotein carrier)*
Canthaxanthin (corrected for lipoprotein carrier)*
30 Lycopene (corrected for lipoprotein carrier)*
Lutein (corrected for lipoprotein carrier)*
anhydro-Lutein (corrected for lipoprotein carrier)*
35 Neurosporene (corrected for lipoprotein carrier)*
Phytofluene (corrected for lipoprotein carrier)*
Phytoene (corrected for lipoprotein
40 carrier)*
α-Tocopherol (corrected for lipoprotein carrier)*
γ-Tocopherol (corrected for lipoprotein carrier)*
45 Retinol*

Retinol binding protein*
Ascorbic acid*
Fe*
K*
Mg*
Total phosphorus*
Inorganic phosphorus*
Se*
Zn*
Ferritin*
Total iron binding capacity*
Fasting glucose*
Urea nitrogen*
Uric acid*
Prealbumin*
Albumin*
Total protein*
Bilirubin*
Thyroid stimulating hormone T3*
Thyroid stimulating hormone T4*
Cotinine
Aflatoxin-albumin adducts
Hepatitis B anti-core antibody (HbcAb)
Hepatitis B surface antigen (GhsAg+)
Candida albicans antibodies
Epstein-Barr virus antibodies
Type 2 Herpes Simples antibodies
Human Papiloma virus antibodies
Heliocobacter pylori antibodies
Estradiol (E2) (adjusted for female cycle)*
Sex hormone binding globulin*
Prolactin (adjusted for female cycle)*
Testosterone (adjusted for female cycle for women)*
Hemoglobin*
Myristic acid (14:0)*
Palmitic acid (16:0)*
Stearic acid (18:0)*
Arachidic acid (20:0)*

17

The biological samples are analyzed to determine the biomarker value for each component in the biological sample for which a biomarker value is desired. It is to be understood that any component that may be found and measured in a biological sample falls within the scope of the present invention. For example, genetic biomarkers which may be measured in a blood

5   sample, as well as the biomarkers that can be measured in any other appropriate biological sample, may also be included.

Since another feature of the present invention is that of identifying new sets of biomarkers useful for predicting disease and death, the biomarker sets may include biomarkers not

10   previously known to have statistical significance for predicting a specific disease or specific cause of death. Thus, since the total number of biomarkers that may be used is substantially unlimited in principle, the actual number of biomarkers used may, in general, be limited only by practical economic and methodological considerations.

15   Since still another feature of the present invention is that of providing a computer-based system for predicting specified biological conditions within a specific time period or age interval in the future, the total number of biomarker values may be limited to only those biomarker values which have statistical significance for predicting a single specified biological condition. Thus, while it is intended that the subject system is typically used as a

20   general purpose tool for predicting and monitoring most, and, eventually, substantially all major types of diseases and underlying causes of death, use of the methodology disclosed herein may also be directed to one disease or cause of death at a time.

After being collected, the biological samples may be analyzed immediately or the samples

25   may be stored for later analysis. Since it is expected that a large number of samples may be collected in a relatively short period of time and under circumstances not conducive to immediate on-site analysis, the samples are preferably stored for later analysis. Because the samples may typically be stored for a substantial period of time, the samples are typically frozen. The samples are to be stored and transported using conditions that preserve the

30   integrity of the samples. Such techniques are described, for example, in Chen, J., Campbell, T. C., Li, J., and Peto, R. Diet, life-style and mortality in China. A Study of the

identified and the underlying cause of death is recorded, preferably using a known coding system, for example, the established International Statistical Classification of Diseases and Related Health Problems, (ICD-10), Geneva, World Health Organization, 1992-c1994, 10th revision. Other coding systems may also be used while remaining within the scope and spirit

5   of the present invention.

Using an effective system to identify when a member of the test population acquires a disease or specified biological condition, morbidity data is also collected, in addition to collecting the biomarker and mortality data of the test population.

10

The database of biomarker values preferably includes information from each individual recording the dates and ages at the times the biomarkers and biomarker samples are collected and recorded, accurate information from the surveillance of the individual recording each incident of disease, medical condition, medical pathology, or death, including diagnosis and

15   date of incident. The database includes values of biomarkers assessed before, during, and after each incident, where feasible.

Since one aspect of the present invention relates to identifying biomarkers not yet known to be statistically significant for predicting future onset of a specified disease or underlying

20   cause of death, as many biomarkers as possible are monitored. In a representative embodiment, about 200 biomarker values are obtained from each member of the test population, although there is substantially no upper limit to the number of biomarkers that may be used to develop the computer-based statistical analysis methodology.

25   Since the present invention is directed toward providing a practical and reliable system for predicting a specified biological condition within a specified period of time or age interval, a substantially complete set of biomarker values is collected from each member of the test population at least two different times. More preferably, so as to obtain information on trends or changes with time, a full set is collected at least three times and, most preferably, the

30   biomarker values are collected at periodic intervals for as long as practically feasible.

21

diagnosed. The time of future onset of the specific health problem occurring for a specific individual can be predicted with a specified quantitative probability estimate based on applying the subject discriminant analysis methodology to the database collected from the large test population. Furthermore, the present invention provides a system for predicting specific health problems further and further into the future with greater and greater reliability as more and more data are collected for ever larger test populations for longer and longer periods of time.

The biological samples are typically analyzed for each biomarker for which quantitative values are desired. For cost and convenience reasons and because of the large number of samples that may be collected, the samples may be analyzed initially only for those individuals already diagnosed with a disease or who die during the time period over which the samples have been collected, as well as for a randomly selected fraction of the remainder of the test population. For example, if the annual mortality rate for the test population surveyed is typically in the range of 2-3% annually, a 300,000 member test population would produce an annual mortality rate of 6000-9,000 deaths, wherein a significant number of deaths would have been caused by each of the major underlying causes of death.

One of the further features of the present invention comprises the step of waiting until a substantial number of deaths have occurred in the test population and then selecting those individuals as the ones for whom the biomarker values are to be determined initially. In addition, a group of still living test members may then be selected from the remainder of the test population. Because of the need to balance the need for large enough numbers of samples to obtain statistically significant results with the need to control costs, the subject system provides a practical method of limiting the analytical measurement costs to only those samples that will tend to provide the most information for the least cost. Naturally, as more and more deaths occur in the test population, larger and larger numbers of samples will be analyzed over time. However, the value of the data obtained, from the point of view of establishing more and more reliable quantitative predictions of future health, will be more or less commensurate with the cost of acquiring the additional biomarker values. This is another of the many special features of the present invention that distinguishes it from any known

23

substantial waste of resources and severe degradation of the quality of the results generated by the remaining data. The subject computer-based methodology includes a feature that provides a means of using substantially all data collected, by using a statistically verifiable technique for filling in the "missing values." This is a particularly useful aspect of the subject

5   methodology, which is based on collecting what amounts to huge quantities of data, as compared with any prior art studies, for very large numbers of test members from a test population that is widely dispersed geographically. Acquisition of comprehensive data from a diverse large test population is particularly desirable so as to obtain biomarker values from members having widely divergent dietary and lifestyle practice representative of the entire

10   human experience.

For the purpose of describing the present invention, the following terminology is explained herein:

15   A "specified biological condition" may, for example, refer to any one of the following:

- a specified disease, for example, as classified in <u>International Statistical Classification of Diseases and Related Health Problems</u>, *supra*. (*e.g.*, diabetes mellitus);

20   - a specified medical or health condition or syndrome (*e.g.*, hypertension, as generally defined by deviations of biomarker or biomarker set values from the usual normal distributions);

- a specified medical event and its sequelae (*e.g.*, ischemic stroke and subsequent death,

25   or non-death and stroke-related partial paralysis and related conditions; myocardial infarction and subsequent death, or non-death and MI-related conditions);

- premature death from any cause (*premature* death at an age earlier than the mean age at death as projected from the person's gender and age at first evaluation);

30

- death at a specified age;

are predicted to acquire the specified biological condition within the specified timeframe. *i.e.*, projected to belong to Group D. These persons are described as having a prescribed *high* probability of acquiring the specified biological condition within the specified timeframe.

5

- Group $\overline{PD}$: That group of persons who, at the beginning of the specified timeframe, are predicted *not* to acquire the specified biological condition within the specified timeframe, *i.e.*, projected to belong to Group $\overline{D}$. These persons are described as having a prescribed *low* probability of acquiring the specified biological condition within a specified timeframe.

10

The term "prescribed high probability" may vary in magnitude from having a probability as low as a few percent, perhaps even as low as 1% or less, or may be as high as 10%, 20%, 50%, or even substantially higher, depending on the specified biological condition. For example, the increased risk of acquiring lung cancer due to smoking may be perceived by many as a significant and preferably avoidable risk, even though the actual several-fold increase in risk that is caused by smoking may *only* be in the range of a 5-10% probability for acquiring lung cancer as far as 15-20 years or more into the future. In any case, for each specified biological condition for which the system is applied, a quantifiably prescribed probability may be determined. The "prescribed low probability" may be specified simply as the probability of not being in the high risk group for acquiring the specified biological condition or, alternatively, the term may be separately specified as a concrete value.

15

20

At the point when a statistically adequate number of the members of the test population can be identified as belonging to Group D or Group $\overline{D}$, the biomarker values of the members of Group D may be compared with members of Group $\overline{D}$ using the subject methodology, so as to determine a statistical procedure for classifying members into Groups PD and $\overline{PD}$ or for estimating the probability, for each member of the test population, of acquiring the specified biological condition within the specified time period or age interval, *i.e.*, the probability of belonging to Group PD or the probability of belonging to Group $\overline{PD}$. In a representative embodiment of the subject invention, the statistical procedure for classifying members into

25

30

27

Phase I. Establish Evaluation Methodology and Select Biomarkers for Consideration.

The following steps would appear in a representative embodiment of the subject invention.

5 *Step 1: Select a methodology for estimating the procedure's error rates.*

The methodology may incorporate any statistically appropriate method of estimating the error rates. Two methods, of many that may be used, are: Training sample/validation sample, and subsampling (or "resampling").

10

Training Sample/Validation Sample Method In the training sample/validation sample approach, the test population is randomly divided into two subsets, identified herein as a "training sample" and a "validation sample. Every subject (member of the test population) is assigned to either the training sample or the validation sample. The data from subjects in the

15 training sample are used in the statistical analyses leading to specification of the discriminant procedure and probability estimation procedure. The data from subjects in the evaluation sample will be used to estimate the discriminant procedure's error rates and the distribution of the probability estimates.

20 *Subsampling Methods* "Subsampling" refers to a class of statistical methods, including jackknifing and bootstrapping, that can be used to produce reduced-bias estimates of error rates. In a subsampling method, data from all subjects are used in the statistical analyses leading to specification of the discriminant procedure and/or distribution of probability estimates. Utilizing all the data can lead to a better discriminant procedure and/or probability

25 estimation procedure than would be obtained in the Training Sample/ Validation Sample approach, especially: (1) if the test population is not large, or, (2) if the *a priori* probability of acquiring the biological condition is small, even with a large test population. In the present context, subsampling methods are computationally intensive.

30 *Step 2. Select the "training sample," i.e., the subset of the test population to be used for statistical analyses leading to the discriminant procedure/probability estimation*

29

Biomarkers will include all recorded, quantitative, personal characteristics of subjects in the test population. The list will include characteristics that do not change over time (e.g., date of birth) as well as time-dependent characteristics, such as body weight or a lab assessment from blood or urine. Non-quantitative characteristics, e.g., the name of the subject's favorite color,
5   will be excluded.

Some of the Potential Biomarkers listed in Step 3 will not be useful for discrimination. The remaining steps of this Phase compile a set of "Candidate Biomarkers," from the Step 3 list of Potential Biomarkers. Each Candidate Biomarker will be selected because there is
10   information from previous research/knowledge, or quantitative evidence from the training sample data, that the biomarker is a potentially useful discriminator. At each step, a biomarker that is selected as a candidate is removed from the list of Potential Biomarkers and moved to the set of Candidate Biomarkers. The reason for removing a selected Candidate Biomarker from the list of Potential Biomarkers: once a biomarker has been selected as a
15   candidate there is no reason to reconsider it; it has already "made the list." At the end of the process, all unselected Potential Biomarkers will be removed from further consideration; only the Candidate Biomarkers will be subjected to additional analyses.

Step 4: Initiate the set of Candidate Biomarkers by including any Potential Biomarkers that,
20   *on the basis of previous research and experience, are confidently believed to be*
     related to the specified biological condition.

The objective of this step is to utilize prior information on biomarkers that are potentially important discriminants for the specified biological condition. For example, if the specified
25   biological condition is acquiring coronary heart disease (CHD) within a specified time, previous research has shown that values of serum cholesterol, systolic blood pressure, glucose intolerance, or cigarette smoking (to name just a few) are related to onset of CHD and should be copied from the list of Potential Biomarkers to the list of Candidate Biomarkers.

30   Any reliable source of information or 'educated guess' may be relied upon to select the subset of biomarkers known or believed to be related to the specified biological condition. Although

31

Step 6: Fit a logistic regression model for each Potential Biomarker, using a binary indicator variable for the specified biological condition as the dependent (Y) variable and age and the Potential Biomarker as the independent (X) variables. Add to the list of Candidate Biomarkers each Potential Biomarker that is "statistically significant"
5 *in its logistic regression model.*

The objective of this step is to select as Candidate Biomarkers those Potential Biomarkers that are related to the probability of acquiring the specified biological condition, after taking the (linear) effect of age into account. The logistic model expresses the probability of
10 acquiring the specified biological condition as a function of the value of the Potential Biomarker, in conjunction with a subject's age.

A biomarker is selected (or not) on the basis of a marginal p-value for the biomarker's slope in the logistic regression model. As with the correlations above, "statistical significance" is
15 used here only as a tool for deciding between "probably important" and "probably unimportant" discriminators. In a representative embodiment, a traditional p-value will be computed for the slope of a Potential Biomarker. If p is less than some specified value, *e.g.*, $p < 0.05$, or $p < 0.01$, the Potential Biomarker is moved to the Candidate Biomarker list.

20 *Step 7: Evaluate each longitudinally-assessed Potential Biomarker, using a general linear mixed model ("MixMod") to assess whether longitudinal trends in the biomarker's values are related to acquisition of the specified biological condition. Each Potential Biomarker with a statistically significant longitudinal trend is moved to the list of Candidate Biomarkers.*
25

The goal of this step is to identify biomarkers, other than those previously promoted to Candidate Biomarker status, that have longitudinal trends that are related to the probability of acquiring the specified biological condition.

30 In a typical embodiment of the subject invention, each model will be created as follows. The dependent variable (Y) in the MixMod contains longitudinal values of the Potential

33

that subject's data are deleted from the analyses.

Given estimates of the mean vectors, μ_i, and covariance matrices, Σ_i, and the biomarker (and related data) for a subject in a vector, Y, the traditional discriminant functions (linear if

5      $\Sigma_1 = \Sigma_2$, quadratic if $\Sigma_1 \neq \Sigma_2$) are evaluated solely from Y, μ_1, μ_2, Σ_1, and Σ_2. The only information specific to the particular subject is in the vector Y.

The mixed model procedure, which is the greater part of Phase II, improves the traditional procedure by using a general linear mixed model (MixMod) to model all of μ_1, μ_2, Σ_1, and

10   Σ_2; the modeled estimates of these parameters are used in the discriminant function rather than the traditional simple, unmodeled estimates. This MixMod procedure makes the following important improvements over traditional discriminant analysis:

- The parameters are estimated using a Mixed Model, that:

15   ◆   uses all available data, i.e., does not use casewise deletion;

◆   supports covariate adjustment of the estimated expected values (μ_i), with corresponding adjustment of the estimated covariance matrices Σ_i, and

◆   supports the utilization of repeated measures (e.g., from annual visits) from the same subject.

20   • This MixMod procedure utilizes model-based estimates of individual random effects and "BLUPs" ("Best Linear Unbiased Predictors"), in addition to or in place of the estimates of the population means μ_i, which can substantially increase the discrimination capability of the discriminant function.

25   *Overview of the Phase II Procedure*

As a result of Phase I, each Candidate Biomarker will have historical or quantitative evidence of utility as a discriminator. However, there are substantial correlations among the Candidate Biomarkers. Consequently, a biomarker that, considered by itself, has substantial

30   discriminatory power, may not make a substantial contribution when used in combination with other biomarkers. In addition, the scales of the biomarkers may vary widely.

"Response—Scaled"). The sample standard deviation of *RespScal* is also approximately 1.00. This scaling facilitates convergence of the iterative procedure in subsequent mixed model computations.

5   Step 1 is executed only once. Initially, all Candidate Biomarkers have data in *RespScal* and are considered members of the set of Select Biomarkers. Non-discriminating biomarkers will be removed from the Select Biomarkers in Steps 2-3.

Step 2: Fit a general linear mixed model (MixMod) using the specifications listed below:

10   *obtain estimates of the parameter matrices* β, Δ, *and* V, *obtain estimates of each subject's random subject effects,* d_{ik} *and each subject's "predicted values," * $Y_{ik}^{(min)}$ *and* $Y_{ik}^{(avg)}$ *as if the subject were in each specified biological condition group,* $i = 1, 2$.

In a representative embodiment of the subject invention the following are specifications of

15   the MixMod:

Dependent (*Y*) variable: *RespScal* ;

Independent (*X*) variables and their coefficients (β):

"Biological Condition Status," an indicator variable for the status of the specified biological condition (classification variable); Biological

20   Condition Status = 1 if the corresponding element of Y contains information about a subject from Group D and Biological Condition Status = 0 otherwise.

Biomarkers' indicator variables (classification variables);

Biological Condition Status × Biomarkers' indicator variables (classification

25   variables);

Age (in years, centered at approximately the overall mean age of subjects; continuous variable);

Random effects variables (Z_k) and random coefficients (effects, d_{ik}):

Subject × Biomarker indicator variables (part of Z_k) and corresponding

30   random effects (intercept increments; part of d_{ik}).

The random subject effect for a specific biomarker is cons... ...ross

37

$d_{k2,\ldots}$]' be the vector of random effects for the k-th subject and h-th scaled biomarker. let $V(d_k) = \Delta = [\delta_{bb'}]$, where $\delta_{bb'} = \mathrm{Cov}(d_{kb}, d_{kb'})$ where h and b' index possibly different scaled biomarkers. let Z_k contain indicator variables for the scaled biomarkers, and let $V_{kb} = \lambda_b I$. Then $\Sigma_k = Z_k \Delta Z'_k + V_k = [\Sigma_{k,bb'}]$, where $\Sigma_{k,bb} = \delta_{bb}J + \lambda_b I =$ covariance matrix of multiple measurements from scaled biomarker b. and $\Sigma_{k,bb'} = \delta_{bb'} \cdot J =$ covariance of scaled biomarkers b and b' evaluated on the same occasion or on different occasions. (Each element of the square matrix J equals 1.)

The process of fitting the mixed model produces estimates of :

> The model's parameters. β, Δ. and parameters of V_k. If the model assumes different covariances for the two Biological Condition Status groups, the model produces separate estimates of the covariance parameters in Δ_i and V_{ik}.

> The expected value of each subject's data vector, μ_{ik}, (subject k being in Biological Condition Status group i),

> The expected value of each subject's data vector, $\mu_{i'k}$, *as if the subject were in the other response group (i')*.

> Each subject's random subject effect in the subject's actual treatment group (i). d_{ik}. and also *as if the subject were in the other response group (i')*, $d_{i'k}$.

> Each subject's "predicted values," in the subject's actual treatment group (i): $Y_{ik}^{(p)}$, and also *as if the subject were in the other response group (i')*: $Y_{i'k}^{(p)}$.

> The subject's covariance matrix, Σ_k. If the model assumes different covariances for the two Biological Condition Status groups, the model produces separate estimates of the covariance matrices Σ_{ik}.

Step 3: Delete the biomarker that has the least apparent discriminant power and re-fit the mixed model.

A biomarker that will be an effective discriminant should have a large (statistically

addition, the covariance parameter matrices Δ_i and V_{ik} may have structure that can be exploited in the analysis, especially when Σ_{ik} is very large, i.e., when there are many biomarkers and/or many longitudinal assessments of one or more biomarkers.

5   The objective of Step 4 is to determine the structure of the covariance parameter matrices Δ_i and V_{ik} for use in the Phase III discriminant analyses. Estimates of large, structured covariance parameter matrices tend to be more precise than estimates of unstructured covariance parameter matrices. A more precise estimate of Δ_i and/or V_{ik} leads to a more precise estimate of $\Sigma_{ik} = Z_{ik}\Delta_i Z'_{ik} + V_{ik}$, thence to more precise estimates of β, the d_{ik}, and

10   the $Y_{ik}^{(p)}$, and to more precise values of the discriminant function.

The overall structure of Σ_{ik} must take into account the following types of covariances/correlations:

Type ADB: Covariances/correlations among different biomarkers evaluated at the

15   same time point;

Type ALESB: Covariances/correlations among longitudinal evaluations of a single biomarker;

Type BTBEL: Covariances/correlations between two biomarkers, evaluated longitudinally, i.e., covariances/correlations between any pair of biomarkers,

20   one evaluated at one time and the other evaluated at a different time.

In a representative embodiment of the subject invention, the structures described in Step 2, above, or extensions of these structures may be useful.

In a representative embodiment of the subject invention, the techniques described in Tangen,

25   Catherine M., and Helms, Ronald W., (1996), "A case study of the analysis of multivariate longitudinal data using mixed (random effects) models," presented at the 1996 Spring Meeting of the International Biometric Society, Eastern North American Region, Richmond, Virginia, March, 1996, are used to explore covariance/ correlation structures for longitudinal multivariate data. Selecting a covariance model typically requires fitting a number of

30   MixMods, typically using the same expected-value model and varying the covariance model. Models may be compared via Log Likelihood statistics (assuming underlying normal

41

A second objective is to estimate the probabilities that a subject will belong to Groups D and D̄.

The technology for achieving the first objective -- classifying a subject into one of the two groups -- uses discriminant procedures that are modifications of traditional discriminant analysis. The estimates of the probability that the subject will be in the group of subjects that will acquire the specified biological condition is obtained from a modification of traditional logistic regression, (1) using the discriminant function values as regressors and (2) using the discriminant variables as regressors.

As noted in the background of Phase II, prior art discriminant analysis methodology typically utilizes naive estimates of the mean vectors, μ_i, and covariance matrices, Σ_i, of the distributions of the biomarkers of the two groups. Moreover, prior art discriminant analysis is typically based upon a "casewise deletion" procedure: if a subject has any missing data, all of that subject's data are deleted from the analyses.

The mixed model procedure, described in Phase II, improves the traditional procedure by using a general linear mixed model (MixMod) to model all of μ_1, μ_2, Σ_1, and Σ_2; the modeled estimates of these parameters are used in the discriminant function rather than the traditional simple, unmodeled estimates. The use of the mixed model permits the present procedures to make the following important improvements over traditional discriminant analysis: The parameters are estimated using all available data, i.e., does not use casewise deletion. The procedure supports covariate adjustment of the estimated expected values (μ_i), with corresponding adjustment of the estimated covariance matrices Σ_i. And the procedure supports the utilization of repeated measures (e.g., from annual visits) from the same subject.

Perhaps more importantly, the use of the mixed model permits the present procedures to utilize model-based estimates of individual random effects and "BLUPs" ("Best Linear Unbiased Predictors"), in addition to or in place of the estimates of the population means μ_i, which can substantially increase the discrimination capability of the discriminant function.

and may or may not include random subject effects.

Phase III Procedure   The steps of Phase III of the procedure are described below. It is
assumed that data are available from one or more "new" subjects, *i.e.*, subjects whose group
5   membership is unknown and that were not used in the Phase II mixed model computations.
In Steps 1-2 we shall consider one subject at a time. Some additional notation is useful. Let i
= 1 for Group D or Group PD, $i=2$ for Group $\bar{D}$ or Group $\overline{PD}$ and let:

Y denote the vector of values of the discriminant variables for one new subject. The
10   elements of Y are scaled as *RespScal* was scaled in Phase II.

X_i denote the matrix of values of the independent variables used in the final Phase II
mixed model, as if the subject were in group i, $i = 1, 2$. Note that the rows of
X_i correspond to the rows (elements) of Y.

Z_i denote the matrix of values of the random effect variables used in the final Phase II
15   mixed model, as if the subject were in group i, $i = 1, 2$. Note that the rows of
Z_i correspond to the rows of Y.

$\hat{\Delta}_i$ denote the estimated covariance matrix of the random effects from group i, $i = 1$,
2, from the final Phase II mixed model. Note that in many cases the mixed
model reduced to a single covariance for the random effects, *i.e.*, $\hat{\Delta}_1 = \hat{\Delta}_2 = \hat{\Delta}$

20

$\hat{V}_i$ denote the estimated covariance matrix of the random residuals or "error terms"
from group i, $i = 1, 2$, from the final Phase II mixed model. Note that in many
cases the mixed model reduced to a single covariance matrix, *i.e.*, $\hat{V}_1 = \hat{V}_2 = \hat{V}$.

25   $\hat{\Sigma}_i = Z_i \hat{\Delta}_i Z_i' + \hat{V}_i$ denote the estimated covariance matrix of Y, from the final Phase
II mixed model, as if the new subject came from group i, $i = 1, 2$. Note that in
many cases the mixed model reduced to a single covariance matrix, *i.e.*, $\hat{\Sigma}_1 = \hat{\Sigma}_2 = \hat{\Sigma}$.

group 1 (Group PD). if $D(Y) \geq 0$; otherwise assign the subject to group 2 (Group $\overline{PD}$).

- If the decision $\Sigma_1 \neq \Sigma_2$ was made in Phase II, evaluate the quadratic discriminant function, $Q(Y)$ (above), substituting $\hat{Y}_i$ for μ_i and $\hat{\Sigma}_i$ for Σ_i, $i = 1, 2$. Assign the subject to group 1 (Group PD) if $Q(Y) \geq 0$; otherwise assign the subject to group 2 (Group $\overline{PD}$).

Classification based upon the "minimum" random subject effects and predicted values, $Y_i^{(min)}$:

- If the decision $\Sigma_1 = \Sigma_2 = \Sigma$ was made in Phase II, evaluate the linear discriminant function, $D(Y)$ (above), substituting $Y_i^{(min)}$ for μ_i and $\hat{\Sigma}$ for Σ. Assign the subject to group 1 (Group PD) if $D(Y) \geq 0$; otherwise assign the subject to group 2 (Group $\overline{PD}$).

- If the decision $\Sigma_1 \neq \Sigma_2$ was made in Phase II, evaluate the quadratic discriminant function, $Q(Y)$ (above), substituting $Y_i^{(min)}$ for μ_i and $\hat{\Sigma}_i$ for Σ_i, $i = 1, 2$. Assign the subject to group 1 (Group PD) if $Q(Y) \geq 0$; otherwise assign the subject to group 2 (Group $\overline{PD}$).

Classification based upon the "average" random subject effects and predicted values, $Y_i^{(avg)}$:

- If the decision $\Sigma_1 = \Sigma_2 = \Sigma$ was made in Phase II, evaluate the linear discriminant function, $D(Y)$ (above), substituting $Y_i^{(avg)}$ for μ_i and $\hat{\Sigma}$ for Σ. Assign the subject to group 1 (Group PD) if $D(Y) \geq 0$; otherwise assign the subject to group 2 (Group $\overline{PD}$).

- If the decision $\Sigma_1 \neq \Sigma_2$ was made in Phase II, evaluate the quadratic discriminant function, $Q(Y)$ (above), substituting $Y_i^{(avg)}$ for μ_i and $\hat{\Sigma}_i$ for Σ_i, $i = 1, 2$. Assign the subject to group 1 (Group PD) if $Q(Y) \geq 0$; otherwise assign the subject to group 2

47

false positive classification, select the procedure with the smaller false negative error

rate, r_{FN}. This situation could arise, for example, if Group D is the subpopulation of

persons who will suffer a myocardial infarction ("MI") within a specified five year

age group. A false negative classification, failure to warn a person of a high MI

5 probability, could have more serious consequences than a false positive classification,

warning a low-probability person that they have a high MI probability.

• Conversely, if a false positive classification has substantially more serious

consequences than a false negative classification, select the procedure with the smaller

10 false positive error rate, r_{FP}.

• When there is no *a priori* reason to assign greater seriousness to either a false negative

or a false positive classification, select the procedure with the smaller total error rate,

r_{tot}.

15 The procedure selected as the apparently most reliable procedure is used to classify subjects

into the two groups, Group PD and Group $\overline{PD}$.

Step 2: Use two types of logistic regression to compute estimates of the probability that a

new subject will belong to each group.

20

The data from the training sample are used to fit a logistic regression model in which the

value of the discriminant function ($D(Y)$ if linear, $Q(Y)$ if quadratic) for each subject is used

as the independent ("X") variable and the Biological Condition Status (indicator variable for

membership in Group D) as the dependent ("Y") variable. The model is used, together with

25 inverse logistic transform, to compute for each subject an estimate of the probability that the

subject will belong to Group D.

In a separate calculation, the data from the training sample are used to fit a logistic regression

model in which the biomarkers used in the discriminant function, together with the final

30 mixed model covariates (variables in X), are incorporated as independent ("X") variables and

the Biological Condition Status (indicator variable for membership in Group D) as the

49

The data used as a basis for this example were obtained from a database including patients for whom Sickle Cell data are acquired on an annual basis. Some patients have data from three consecutive visits. However, since patients typically cannot be compelled to participate annually, the database includes many patients for whom data are available from only one or

5   two annual visits. Database information that was used here included demographic data, clinical chemistry data, and hematological data.

The specified biological condition of interest (the "disease" or "affliction") in this example was an occurrence of a painful crisis that required hospitalization. At each annual visit the

10   subject is asked (and records are checked to determine) if the subject had a painful crisis that required hospitalization in the preceding year. Each subject who reported having had a hospitalization for a painful crisis at any visit (any year) is a member of the "Diseased" group (Group D); all other subjects are members of Group $\bar{D}$.

15   Whenever a subject had had a painful crisis that required hospitalization in the preceding year, all data that were collected after the hospitalization for the painful crisis, in the same year or in later years, were excluded from the analysis. This mimics the procedure that would be used if the outcome were death or occurrence of a chronic, incurable disease. The variable that records a subject's Group D membership (e.g., diseased or not, afflicted or not) is named

20   the "Disease Status" variable.

The following is an example of the statistical analysis procedures using the sickle cell data. For reasons of confidentiality, the data used in this example are artificial and do not come from a real study or from real subjects. However, the data are similar to data that could have

25   been obtained in a study of real subjects.

Step 5: *Add to the list of Candidate Biomarkers any Potential Biomarkers that are*
"statistically significantly" correlated with the "known important" biomarkers from
Step 4.

5   Biomarkers were selected that were correlated with the "known important" biomarker,

platelets, from Step 2. A summary of these correlations is shown in Table 3, in the columns

labeled "Correlation W/ Platelets". The "p" column shows the p-values for correlations with

Platelets. A biomarker was selected on the basis of a marginal p-value for the Pearson

product-moment correlation coefficient. In the example, $p < 0.01$ was required for selection.

10   The "$p<cv$" column indicates, by the presence of the word "YES," those biomarkers that

became Candidate Biomarkers as a result of a "significant" correlation with Platelets.

Step 6: *Fit a logistic regression model for each Potential Biomarker, using a binary*
indicator variable for the specified biological condition as the dependent (Y) variable
15   *and age and the Potential Biomarker as the independent (X) variables. Add to the list*
of Candidate Biomarkers each Potential Biomarker that is "statistically significant"
in its logistic regression model.

A logistic regression model was fitted for each biomarker, using Disease Status as the

20   dependent (Y) variable and a combination of age and the biomarker as the independent (X)

variables. In this case, for each biomarker the logistic model assessed how well the

probability of a hospitalization for a painful crisis is described by that biomarker, in

conjunction with the subject's age. Roughly speaking, the biomarker's regression coefficient,

or slope, in the logistic regression will be approximately zero if there is no relationship

25   between the biomarker and the probability that the subject will acquire the specified

biological condition; a nonzero slope indicates a relationship. A summary of the logistic

regression results is shown in Table 3, in the columns headed "Logistic Regression." The "p"

column shows the p-values for the biomarker's regression coefficient. A biomarker was

selected on the basis of a marginal p-value for the biomarker's slope in the logistic regression

30   model. In the example, $p < 0.01$ was required for selection. The "$p<cv$" column indicates,

by the presence of the word "YES," those biomarkers that became Candidate Biomarkers as a

At the end of Steps 4-7. all Potential Biomarkers have been examined and each biomarker with historical or quantitative evidence of utility as a discriminator has been moved to the list of Candidate Biomarkers. The Candidate Biomarkers are indicated by the word "YES" in Table 3 in the column headed "Selected."

5

Phase II.   Reduce the Candidate Biomarkers to a Set of Select Biomarkers that have Discriminatory Power and Perform Mixed Model Estimation of the Covariance Structure and Predicted Values.

10   *Step 1:   Prepare a dataset in which one variable, "RespScal," contains scaled values (including longitudinal measures) of all Candidate Biomarkers from all subjects.*

This step was executed for the example but the results are not shown. However, note that when all the values of all the different biomarkers are placed into one column vector, Y, the

15   vector can contain a large number of elements.

Step 2:   Fit a general linear mixed model (MixMod) using the specifications listed below: obtain estimates of the parameter matrices β, Δ, and V, obtain estimates of each subject's random subject effects, d_{ik}, and each subject's "predicted values," $Y_{ik}^{(min)}$

20   *and $Y_{ik}^{(avg)}$ as if the subject were in each specified biological condition group, i=1, 2.*

Step 3:   Delete the biomarker that has the least apparent discriminant power and re-fit the mixed model.

25

Steps 2-3 are repeated iteratively until all biomarkers in the model are statistically significant. In the interests of conserving space in this presentation of an example, only the final results of the iterations through Steps 2-3 are discussed. Steps 2-3 reduced the number of biomarkers to 15, with Age as a fixed effect covariate.

30

General information for the example mixed model is given in Table 4. Data were available

55

one more biomarker.

Step 4: Determine the structures of the covariance parameter matrices, Δ_r and V_{ik}.

As noted above, the overall structure of Σ_{ik} must take into account three types of covariances/correlations:

Type ADB: Covariances/correlations among different biomarkers evaluated at the same time point;

Type ALESB: Covariances/correlations among longitudinal evaluations of a single biomarker;

Type BTBEL: Covariances/correlations between two biomarkers, evaluated longitudinally, *i.e.*, covariances/correlations between any pair of biomarkers, one evaluated at one time and the other evaluated at a different time.

In the example the following structures were ultimately obtained:

Identical random effects covariance parameter matrices for both Group D and Group D̄, *i.e.*, $\Delta_1 = \Delta_2 = \Delta$ and

Δ has compound symmetric structure, $\delta_{ii} = 0.6669$, $\delta_{ij} = 0.0097$ for $i \neq j$.

Type ADB covariances in matrix V, which is the same for both Group D and Group D̄, and compound symmetric structure, $v_{ii}=0.3267$, $v_{ij}=0.0151$ for $i \neq j$.

This covariance structure was reasonable given the sickle cell data at hand.

Estimates of Δ and V are shown in Table 7. The estimate of Δ, the covariance matrix of the random subject effects, is in the top of the table. The rows and columns correspond to the 15 biomarkers used in this model; the columns are labeled.

The estimate of V, the covariance matrix of the within-subject, within-visit errors, is in the bottom of the table. As with Δ, the rows and columns correspond to the 15 biomarkers used in this model. V has compound symmetric structure, which is reasonable for the scaled data.

Phase III: Calculate Discriminant Functions Using Estimated Means and Predicted

rates than are likely to occur in practice, because the training sample was used both to derive the discriminant function and to evaluate it. Evaluation of the discriminant function using the evaluation sample will produce unbiased estimates of the misclassification rates. Resampling techniques such as jackknifing or bootstrapping can produce less biased estimates while still

5   using data from the training sample.

Step 2:   Use two types of logistic regression to compute estimates of the probability that a new subject will belong to each group.

10

Two types of logistic regressions are fitted to the training sample data for each of the discriminant functions. In both logistic regressions, the Disease Status indicator is the dependent ("I") variable. In the first logistic regression, the value of the discriminant functions based on estimation is used as an independent ("X") variable. In the second

15   logistic regression, the value of the discriminant functions based on prediction is used as an independent ("X") variable. In a third logistic regression, the biomarkers used in the discriminant function are incorporated as independent ("X") variables, along with covariates used in the fixed effects part of the mixed model, and the Disease Status indicator is the dependent ("I") variable. The estimates from the logistic regression models are used to

20   compute, for each subject, an estimated probability that the subject belongs to the diseased (Disease Status "Yes") group. The results of the logistic regression computations are not displayed in tables.

Figure 1 displays the empirical distribution functions ("EDF") of the linear discriminant

25   function values (based on estimated values) for Group D (solid line) and Group $\bar{D}$ (dashed line). To prepare the graph, the data for the subjects are sorted by Disease Status group and, within a group, by increasing values of $D(Y)$. Data points are plotted in that sequence. The EDF value starts at 0 (before the first subject's data are plotted) and increases by $1/n$ for each subject, where n is the number of subjects in that group. Thus, the EDF climbs from 0 to 1,

30   separately for each group. In Figure 1, the fact that the EDF for Group D is shifted to the left of the EDF for Group $\bar{D}$ indicates that Group D tends to have lower scores than Group $\bar{D}$.

Table 2. Description of Potential Biomarkers for the Sickle Cell Data

Variable Name	Description
AGEYR	Age of patient (years)
ALBUMIN	Albumin (g/dL)
ALKPHOS	Alkaline Phosphatase (u/L)
BMI	Body Mass Index (Wt./ Ht.2)
BP_DIAST	Diastolic Blood Pressure (mm Hg)
BP_SYST	Systolic Blood Pressure (mm Hg)
CALCIUM	Calcium (g/dL)
CL	Chloride (meq/L)
CO2	Carbon Dioxide (mmol/L)
GENDER	Gender of patient (M/F)
HBA2	Hemoglobin A2 (%)
HCT	Hematocrit (%)
HEIGHT	Height (cm)
HGB	Hemoglobin (g/dl)
K	Potassium (mmol/L)
L_ALKPH	Log10 of Alkaline Phosphatase
L_ALT	Log10 of Alanine Transaminase
L_AST	Log10 of Aspartate Transaminase
L_BUN	Log10 of Blood Urea Nitrogen
L_CR	Log10 of Creatinine
L_DBILI	Log10 of Direct Bilirubin
L_HBF	Log10 of Hemoglobin F
L_LDH	Log10 of Lactic Dehydrogenase
L_TBILI	Log10 of Total Bilirubin
L_URICA	Log10 of Uric Acid
MCH	Mean Corpuscular Hemoglobin (mg/dL)
MCHC	Mean Corpuscular Hemoglobin Concentration (b/dL)
MCV	Mean Corpuscular Volume (fl)
NA	Sodium (meq/L)
PHOSPHOR	Phosphorus (mg/dL)
PLATELET	Platelet Count (x 109/L)
RBC	Red Blood Cell Count (x 109/L)
RETIC	Reticulocyte Count (%)
TOTPROT	Total Blood Protein (g/L)
WBC	White Blood Cell Count (x 109/L)
WEIGHT	Weight of patient (kg)

Analyte		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(p)	p<cv	(p)	p<cv	(p)	p<cv	(p)	p<cv
L_HBF		95	113	0.83	0.45	80	118	0.75	0.45	0.84		0.11		0.32		0.08	
L_LDH		136	158	2.59	0.32	96	142	2.58	0.24	0.02		0.46		0.48		0.63	
L_TBILI	YES	234	280	0.33	0.31	159	252	0.31	0.31	0.00	YES	0.11		0.48		0.09	
L_URICA		231	278	0.69	0.15	158	250	0.70	0.15	0.83		0.04		0.63		0.25	
MCH	YES	162	195	29.10	4.03	145	219	28.22	4.00	0.00	YES	0.20		0.43		0.13	
MCHC	YES	168	205	33.59	1.30	150	226	34.19	1.60	0.91		0.00	YES	0.62	YES	0.01	YES
MCV	YES	203	250	87.29	11.67	166	259	82.72	11.31	0.00	YES	0.00	YES	0.35	YES	0.00	YES
NA		246	291	140.16	2.76	157	243	140.27	2.32	0.09		0.47		0.14		0.86	
PHOSPHOR	YES	214	257	4.18	0.82	151	236	4.44	0.87	0.01	YES	0.04		0.57		0.04	
PLATELET	YES	199	248	419.49	183.43	168	263	355.79	146.53	1.00		0.00	YES	0.95	YES	0.00	YES
RBC	YES	202	250	3.18	0.92	165	259	3.48	1.01	0.00	YES	0.00	YES	0.18	YES	0.00	YES
RETIC	YES	92	115	8.16	5.64	86	130	6.45	4.91	0.00	YES	0.01	YES	0.56	YES	0.01	YES
TOTPROT	YES	232	279	7.65	0.71	159	251	7.60	0.60	0.00	YES	0.41		0.16		0.96	
WBC	YES	202	251	11.91	4.54	168	263	10.27	4.14	0.00	YES	0.00	YES	0.92	YES	0.00	YES
WEIGHT		263	316	49.44	27.97	175	269	44.04	28.43	0.14		0.52		0.75		0.30	

NOTES for Table 2:

"p<cv" is an abbreviation for "The p-value is less than the critical value; here, cv=0.01.

In "p<cv" columns, a blank means "NO."

All p-values have been rounded to 2 decimal places.

63

Table 4. Mixed Model Information: Overall Model Characteristics

Covariance Parameters	4
Columns in X	31
Max Cols in Z Per Subject	15
Subjects	481
Max Obs Per Subject	90
Observations Used	7254
Observations Not Used	11946
Total Observations	19200

Biomarker Main Effect or Interaction (IA)	$\hat{\beta}$	s.e.$(\hat{\beta})$	t	p-Value
L_BUN X GROUP IA	-0.290	0.101	-2.87	0.0041
MCH X GROUP IA	0.225	0.108	2.08	0.0377
MCHC X GROUP IA	-0.322	0.107	-3.02	0.0025
MCV X GROUP IA	0.390	0.099	3.92	0.0001
PHOSPHOR X GROUP IA	-0.335	0.101	-3.33	0.0009
PLATELET X GROUP IA	0.320	0.099	3.21	0.0013
RBC X GROUP IA	-0.337	0.099	-3.39	0.0007
RETIC X GROUP IA	0.390	0.142	2.75	0.0059
WBC X GROUP IA	0.367	0.099	3.70	0.0002
AGE (CENTERED)	-0.001	0.001	-0.87	0.3848

447	L_BUN	3	NO	4.422	REAL	4.104	4.059
447	MCH	1	NO	7.156	REAL	6.931	7.147
447	MCH	2	NO		REAL	6.930	7.146
447	MCH	3	NO	7.279	REAL	6.929	7.145
447	MCHC	1	NO	21.994	REAL	22.549	22.191
447	MCHC	2	NO		REAL	22.548	22.191
447	MCHC	3	NO	22.259	REAL	22.547	22.190
447	MCV	1	NO	7.438	REAL	7.048	7.399
447	MCV	2	NO	7.378	REAL	7.047	7.398
447	MCV	3	NO	7.600	REAL	7.046	7.397
447	PHOSPHOR	1	NO	3.180	REAL	4.894	3.620
447	PHOSPHOR	2	NO	3.399	REAL	4.893	3.619
447	PHOSPHOR	3	NO	3.728	REAL	4.892	3.618
447	PLATELET	1	NO	2.335	REAL	2.148	2.267
447	PLATELET	2	NO	2.501	REAL	2.147	2.266
447	PLATELET	3	NO	2.073	REAL	2.146	2.265
447	RBC	1	NO	4.688	REAL	3.316	4.511
447	RBC	2	NO	4.593	REAL	3.315	4.510
447	RBC	3	NO	4.871	REAL	3.314	4.509
447	RETIC	1	NO		REAL	1.145	1.197
447	RETIC	2	NO		REAL	1.145	1.196
447	RETIC	3	NO		REAL	1.144	1.195
447	WBC	1	NO	3.069	REAL	2.317	2.561
447	WBC	2	NO	1.873	REAL	2.316	2.561
447	WBC	3	NO	2.911	REAL	2.315	2.560
447	ALBUMIN	1	YES		UNREAL	10.049	9.990

447	MCV	1	YES	UNREAL	7.048	7.399
447	MCV	2	YES	UNREAL	7.047	7.398
447	MCV	3	YES	UNREAL	7.046	7.397
447	PHOSPHOR	1	YES	UNREAL	4.894	3.620
447	PHOSPHOR	2	YES	UNREAL	4.893	3.619
447	PHOSPHOR	3	YES	UNREAL	4.892	3.618
447	PLATELET	1	YES	UNREAL	2.148	2.267
447	PLATELET	2	YES	UNREAL	2.147	2.266
447	PLATELET	3	YES	UNREAL	2.146	2.265
447	RBC	1	YES	UNREAL	3.316	4.511
447	RBC	2	YES	UNREAL	3.315	4.510
447	RBC	3	YES	UNREAL	3.314	4.509
447	RETIC	1	YES	UNREAL	1.145	1.197
447	RETIC	2	YES	UNREAL	1.145	1.196
447	RETIC	3	YES	UNREAL	1.144	1.195
447	WBC	1	YES	UNREAL	2.317	2.561
447	WBC	2	YES	UNREAL	2.316	2.561
447	WBC	3	YES	UNREAL	2.315	2.560

MCHC	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151
MCV	0.0151	0.0151	0.0151	0.0151	0.0151	0.3267	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151
PHOSPHOR	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.3267	0.0151	0.0151	0.0151	0.0151	0.0151
PLATELET	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.3267	0.0151	0.0151	0.0151	0.0151
RBC	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.3267	0.0151	0.0151	0.0151
RETIC	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.3267	0.0151
WBC	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.0151	0.3267

Table 8. Evaluation of the Discriminant Procedure Using Estimated Values

Numbers of subjects in the validation sample tabulated by actual and classified membership in D.		Subject was classified as a member of Group:	
		$\overline{PD}$ No	PD Yes
Subject was actually a member of Group:	$\overline{D}$ No	$N_{11}=100$ $r_{11}=56\%$	$N_{12}=79$ $r_{12}=r_{FP}=44\%$
	D Yes	$N_{21}=74$ $r_{21}=r_{FN}=28\%$	$N_{22}=188$ $r_{22}=72\%$

$r_{tot} = 153/441 = 35\%$

75

What Is Claimed Is:

1. A computer-based system for predicting future health of individuals comprising:

(a) a computer comprising a processor containing a database of longitudinally-acquired biomarker values from individual members of a test population. subpopulation D of said members being identified as having acquired a specified biological condition within a specified time period or age interval and a subpopulation $\bar{D}$ being identified as not having acquired the specified biological condition within the specified time period or age interval; and

(b) a computer program that incudes steps for:

(1) selecting from said biomarkers a subset of biomarkers for discriminating between members belonging to the subpopulations D and $\bar{D}$, wherein the subset of biomarkers is selected based on distributions of the biomarker values of the individual members of the test population; and

(2) using the distributions of the selected biomarkers to develop a statistical procedure that is capable of being used for:

(i) classifying members of the test population as belonging within a subpopulation PD having a prescribed high probability of acquiring the specified biological condition within the specified time period or age interval or as belonging within a subpopulation $\overline{PD}$ having a prescribed low probability of acquiring the specified biological condition within the specified time period or age interval; or

(ii) estimating quantitatively, for each member of the test population, the probability of acquiring the specified biological condition within the specified time period or age interval.

2. The computer-based system of claim 1 wherein the statistical procedure comprises a discriminant function utilizing the estimated mean vectors and estimated covariance matrices of the distributions of biomarker values within the subpopulations D and $\bar{D}$.

3. The computer-based system of claim 2 wherein estimates of parameters of the distributions of the selected biomarkers are obtained by fitting a general linear mixed model to the biomarker

77

selected based on distributions of the biomarker values of the individual members of the test population; and

(2) using the distributions of the selected biomarkers to develop a statistical procedure that is capable of being used for:

(i) classifying members of the test population as belonging within a subpopulation PD having a prescribed high probability of acquiring the specified biological condition within the specified time period or age interval or as belonging within a subpopulation $\overline{PD}$ having a prescribed low probability of acquiring the specified biological condition within the specified time period or age interval; or

(ii) estimating quantitatively, for each member of the test population, the probability of acquiring the specified biological condition within the specified time period or age interval;

wherein the statistical procedure comprises a discriminant function utilizing the estimated mean vectors and estimated covariance matrices of the distributions of biomarker values within the subpopulations D and $\hat{D}$.

9. The computer-based system of claim 8 wherein estimates of parameters of the distributions of the selected biomarkers are obtained by fitting a general linear mixed model to the biomarker data from the test population.

10. The computer-based system of claim 9 wherein:

(a) the estimated mean vectors are modeled as vector-valued functions of expected-value parameters or values of covariates; or

(b) estimated covariance matrices are modeled as matrix-valued functions of covariance parameters or values of covariates.

11. The computer-based system of claim 10 wherein an estimated mean vector or probability incorporates an estimate of the realized value of a random subject effect vector for a member being classified or of a member for whom a probability is estimated.

15. A computer-based system for predicting an individual's future health comprising:

(a) a computer comprising a processor containing a plurality of biomarker values from an individual; and

(b) a computer program that incudes steps for applying a statistical procedure to said plurality of biomarker values so as:

(i) to classify said individual as having a prescribed high probability of acquiring a specified biological condition within a specified time period or age interval or as having a prescribed low probability of acquiring the specified biological condition within the specified time period or age interval; or

(ii) to estimate quantitatively for said individual the probability of acquiring the specified biological condition within the specified time period or age interval;

wherein said statistical procedure is based on :

(1) collecting a database of longitudinally-acquired biomarker values from individual members of a test population, subpopulation D of said members being identified as having acquired the specified biological condition within the specified time period or age interval and a subpopulation $\bar{D}$ being identified as not having acquired the specified biological condition within the specified time period or age interval;

(2) selecting from said biomarkers a subset of biomarkers for discriminating between members belonging to the subpopulations D and $\bar{D}$, wherein the subset of biomarkers is selected based on distributions of the biomarker values of the individual members of the test population; and

(3) using the distributions of the selected biomarkers to develop said statistical procedure.

16. The computer-based system of claim 15 wherein the plurality of biomarker values from said individual includes longitudinally-acquired biomarker values.

17. The computer-based system of claim 15 wherein the specified biological condition is death due to a specified underlying cause of death within the specified time period or age interval.

81

underlying cause of death comprising:

(a) a computer comprising a processor containing a plurality of biomarker values from an individual; and

(b) a computer program that incudes steps for applying a statistical procedure to said

5　plurality of biomarker values so as to determine whether said individual is classified as having a prescribed high probability of dying, within a specified time period or age interval, from any one of the underlying causes of death that account in the aggregate for at least 60% of all deaths in a test population over the specified time period or age interval.

10　24. A computer-based system for assessing an individual's evidence of good health comprising:

(a) a computer comprising a processor containing a plurality of biomarker values from an individual; and

(b) a computer program that incudes steps for applying a statistical procedure to said plurality of biomarker values so as to determine whether said individual is classified as having a

15　prescribed high probability of not dying, within a specified time period or age interval, from any one of the underlying causes of death that account in the aggregate for at least 60% of all deaths in a test population over the specified time period or age interval.

25. An apparatus for assessing an individual's risk of future health problems comprising:

20　(a) a storage device for storing a plurality of biomarker values from an individual; and

(b) a processor coupled to the storage device and programmed:

1) to receive from the storage device said plurality of biomarker values; and

2) to apply a statistical procedure to said plurality of biomarker values so as:

(i) to classify said individual as belonging within a subpopulation PD

25　having a prescribed high probability of acquiring a specified biological condition within a specified time period or age interval or as belonging within a subpopulation $\overline{PD}$ having a prescribed low probability of acquiring the specified biological condition within the specified time period or age interval; or

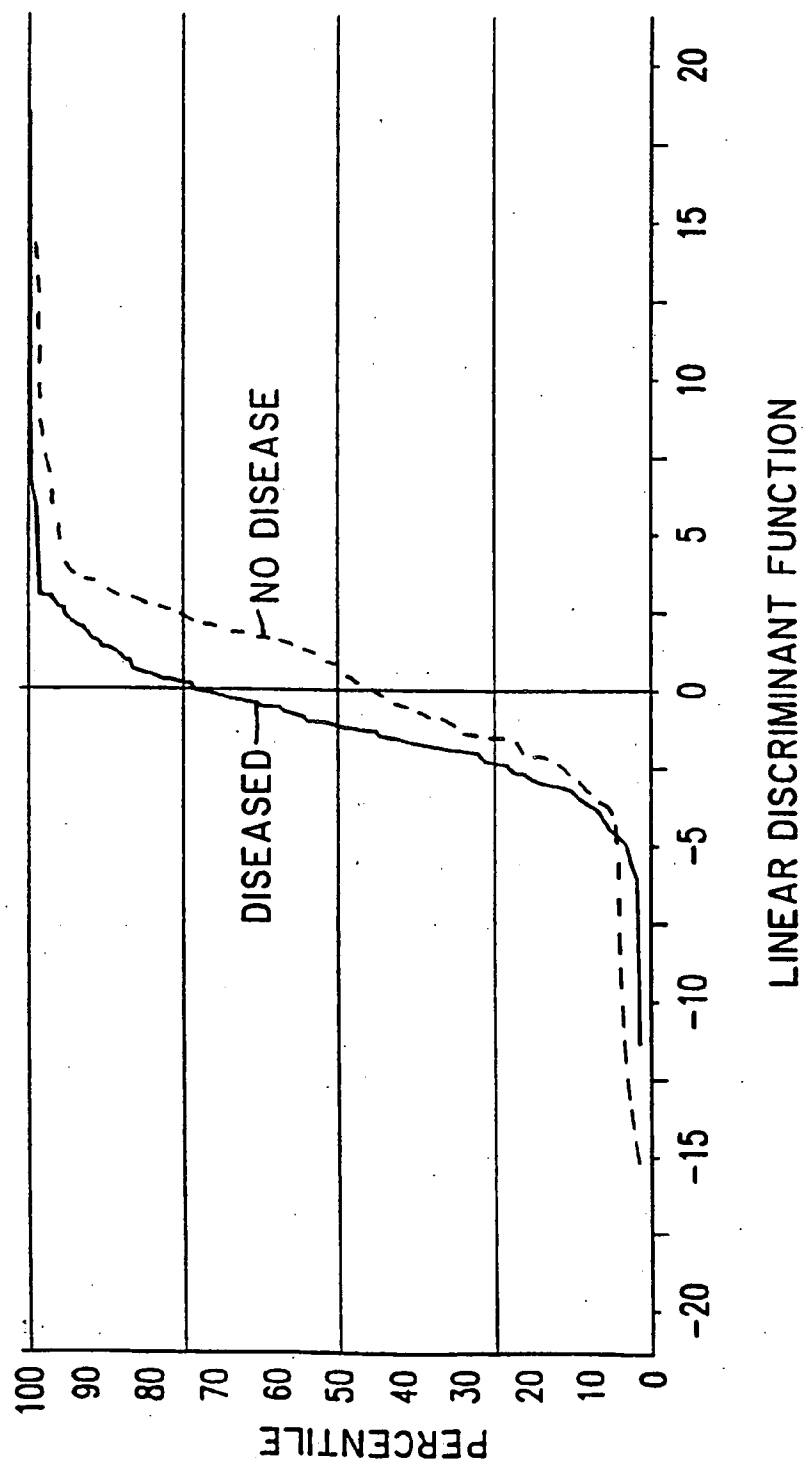(ii) to estimate quantitatively the probability for said individual acquiring

83

FIG. 1

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6)   :A61B 5/00
US CL   :600/300
According to International Patent Classification (IPC) or to both national classification and IPC

B.   FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. :   128/920, 923-925; 600/300, 301

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C.   DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y,P	US 5,687,716 A (KAUFMANN et al) 18 November 1997, entire document.	1-25
Y,P	US 5,629,501 A (MINTURN) 02 December 1997, entire document.	1-25

☐   Further documents are listed in the continuation of Box C.   ☐   See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 MARCH 1998	**1 7 APR 1998**
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	Authorized officer ERIC F. WINAKUR
Facsimile No. (703) 305-3230	Telephone No. (703) 308-3940

Form PCT/ISA/210 (second sheet)(July 1992)*